PB-241 230

A REVIEW OF CLUSTERING TECHNIQUES WITH EMPHASIS ON BENTHIC ECOLOGY

John D. Walker

Pacific Northwest Environmental Research Laboratory
Newport, Oregon

August 1974

DISTRIBUTED BY:

NTIS

National Technical Information Service
U. S. DEPARTMENT OF COMMERCE

YL 120041

PB241230

A Review of Clustering Techniques With Emphasis on Benthic Ecology

by John Walker

Environmental Protection Agency - Newport, Oregon

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
US Department of Commerce
Springfield, VA. 22151

Resources Development Internship Program
Western Interstate Commission for Higher Education

Urban Studies Center,
Portland State University

BIBLIOGRAPHIC DATA SHEET	1. Report No.	2.	PB 241 230

4. Title and Subtitle A review of clustering techniques with emphasis on benthic ecology.	5. Report Date August 1974 6.
7. Author(s) John D. Walker.	8. Performing Organization Rept. No.
9. Performing Organization Name and Address Portland State University, Urban Studies Center, Western Interstate Commission for Higher Education, Resources Development Internship Program, Portland, Oregon	10. Project/Task/Work Unit No. 11. Contract/Grant No.
12. Sponsoring Organization Name and Address Environmental Protection Agency, Marine Science Center, Coastal Pollution Branch, Newport, Oregon 97365	13. Type of Report & Period Covered 14.

15. Supplementary Notes

16. Abstracts

Numerical clustering techniques are reviewed including Agglomerative Hierarchical, Divisive Hierarchical and Non-Hierarchical strategies. The use of these various techniques in Marine Benthic Ecology is also reviewed.

17. Key Words and Document Analysis. 17a. Descriptors

numerical clustering techniques
Marine Benthic Ecology

17b. Identifiers/Open-Ended Terms

Reproduced by
NATIONAL TECHNICAL INFORMATION SERVICE
US Department of Commerce
Springfield, VA. 22151

17c. COSATI Field/Group

PRICES SUBJECT TO CHANGE

18. Availability Statement Available from NTIS	19. Security Class (This Report) UNCLASSIFIED	21. No. of Pages 26
	20. Security Class (This Page) UNCLASSIFIED	22. Price $3.75-2.25

A Review of Clustering Techniques with Emphasis on Benthic Ecology.

By

John D. Walker
Western Interstate Council for Higher Education Intern
Coastal Pollution Branch
U.S. Environmental Protection Agency
Marine Science Center
Newport, Oregon 97365

August 1974

Committee Members:

R. C. Swartz
D. T. Martin
W. A. Deben
D. J. Baumgartner

ABSTRACT

Numerical clustering techniques are reviewed including Agglomerative Hierarchical, Divisive Hierarchical and Non-Hierarchical strategies. The use of these various techniques in Marine Benthic Ecology is also reviewed.

This report has been cataloged by the WICHE Library as follows:

Walker, John D
A review of clustering techniques with emphasis on benthic ecology. Boulder, Colo., Western Interstate Commission for Higher Education, 1974.
(25)p.

1. Marine ecology - Mathematical techniques. I. Title. II. Northwest Pacific Environmental Research Laboratory. III. Western Interstate Commission for Higher Education. Resources Development Internship Program.

The ideas and opinions expressed in this report are those of the author. They do not necessarily reflect the views of the WICHE Commissioners or WICHE staff.

The Resources Development Internship Program has been financed during 1974 by grants from the Economic Development Administration, Education Division, HEW; Jessie Smith Noyes Foundation, National Endowment for the Humanities, National Science Foundation and by more than one hundred and fifty community agencies throughout the West.

WICHE is an Equal Opportunity Employer.

Printed on recycled paper

INTRODUCTION

Numerical grouping techniques have been used in a variety of scientific fields, especially numerical taxonomy. In the past decade, ecologists have employed these methods to elucidate community structure. Because of this diversity of interests, these techniques have been referred to as taxometric methods (Sneath and Sokal, 1973), classification methods (Pielou, 1969), and clustering techniques (Williams, 1971). Clustering techniques seems the most general term because it does not imply a specific use.

In benthic studies, the raw data consists of a number of species taken from several collection sites. Clustering techniques analyze this raw data giving an investigator some idea of the similarities within the entire group. Benthic studies using clustering techniques usually analyze the data two ways: collection sites are considered as the individuals with the species from the sites acting as attributes resulting in site groups; or species are considered as individuals with collection sites as attributes resulting in species groups.

Because of the ability of clustering techniques to sort out community types, it has become of interest in pollution studies, as was done recently by Stephenson et al. (1974) and Roback et al. (1969). Because of pollutional stresses, benthic communities affected by pollution will be sorted from community types in healthy areas, thus facilitating detection of pollution stresses and management of pollution abatement efforts.

Clustering techniques can be divided into three main subgroups, two of which are hierarchical and the third non-hierarchical or reticulate.

The two hierarchical strategies are referred to as agglomerative and divisive. Both strategies define a method of producing clusters of similar individuals from the entire population. Agglomerative hierarchies start with the individuals and fuse groups up to the entire population while divisive hierarchies start with the population and divide it down to the individuals. Non-hierarchical strategies do not define a route between groups but optimize the structure of the group by making it as homogeneous as possible.

Of the two hierarchical strategies, agglomerative seems to have more inherent problems than divisive; however, in marine ecology, the agglomerative hierarchical strategies seem to be more popular. The two main problems with agglomerative strategies are: (1) since a user's interest is normally concentrated in the higher levels of the hierarchy the fusion process must run from the individual to the entire population. For a population of N individuals the fusions necessary equal $(N-1)^2$ (Williams, 1971). For a large N, computation time can be prohibitive. (2) There is a tendency for minor misclassification because the fusion starts at the individual level where chance anomalies are more likely to occur.

Divisive strategies are not as sensitive to these two problems. Since fission starts at the population, the chance anomalous behavior of the individuals is more likely overlooked. Also, starting with the entire population a "stopping rule" can easily be programmed to halt fission at whatever level the investigators interest lies, hence computation time can be reduced considerably. Most divisive strategies are based on MONOTHETIC division (i.e., based on a single attribute, which in the case of benthic studies would be presence-absence) and selecting the right attribute requires considerable insight since it must divide the population into two groups as unlike as possible. Polythetic divisive strategies which divide the population based on more than one attribute

(such as % similarity or dissimilarity) do exist; however, they require a much greater computational time than agglomerative strategies of comparable size.

In this paper, the two hierarchical strategies will be discussed in detail including benthic studies where these methods were employed. Non-hierarchical strategies will also be discussed although they have not received the attention that hierarchical systems have.

AGGLOMERATIVE HIERARCHICAL STRATEGIES (AHS)

Interindividual Measures To initiate the fusion in an agglomerative strategy, there needs to be some measure for comparing individuals. They are concerned with numerical definition of likeness and have been referred to as interindividual measures by Williams (1971). There have been large numbers of these measures proposed (reviewed by Bergen, 1971; Goodman and Kruskal, 1959; and Sokal and Sneath, 1963), all of which fall into three main classes (Williams, 1971). These are (a) Manhattan metric of the basic form $\Sigma|X_{1j} - X_{2j}|$, (b) Euclidean distance $(\Sigma|X_{1j} - X_{2j}|^2)^{\frac{1}{2}}$ (In both cases for site group classification X_{1j} and X_{2j} would be the number of individuals of species (j) in sites 1 and 2), and (c) various forms of information statistics usually using the formulation of Shannon (the specific formulation varies, examples of which will be given later).

Williams (1971) outlined several decisions that must be made prior to clustering. Double-zero matches (which in the case of site classification would mean the species in question was absent from both sites) which are quite common in species rich marine surveys are of particular interest. If double-zero matches are counted toward likeness, the interindividual measure is symetrical; if they are not, then it is asymetrical (Williams, 1971). Both symetrical and asymetrical models are available so the

choice is left to the individual user. Since double-zero matches are not a specific problem of AHS, further discussion of it as well as transformation of attributes and standardization will be discussed in the last section of the paper.

Individual/Group and Group/Group Measure Every AHS begins with the calculation of all interindividual measures. As soon as fusion begins, individual/group and group/group measures are needed. It is desirable that the individual/group and group/group measures are similar to the interindividual measure for then both the interindividual measure and individual/group measure can be considered a special case of the group/group measure. Some of the measures contain what is described by Lance and Williams (1967b) as Combinatorial solutions. In such cases, the individual/group and group/group measures can be calculated directly from a matrix of interindividual measures. This has the advantage that once the interindividual measures are calculated, the raw data is no longer needed for subsequent calculations.

Lance and Williams (1967b) have further shown that the majority of the combinatorial solutions can be encompassed within a single generalized linear model: $d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma|d_{hi} - d_{hj}|$. In this model, d_{hi}, d_{hj} and d_{ij} are all dissimilarity type measures for the individuals of groups h, i and j. Groups i and j are fusing to form group k for which the equation calculates the dissimilarity between h and k (d_{hk}). Values for α_i, α_j, β and γ vary depending on the strategy.

The following strategies are mostly those given by Lance and Williams (1967b). Those with combinatorial solutions which can be calculated by the linear model above will have the appropriate values for α_i, α_j, β and γ.

a. Nearest-neighbor. The distance between the two groups is defined as the distance between their closest individual, one in each group ($\alpha_i = \alpha_j = +\frac{1}{2}$, $\beta = 0$, $\gamma = -\frac{1}{2}$).

b. Furthest-neighbor. It is the opposite of Nearest-neighbor in that the distance between two groups is defined as that between the most remote pair of elements, one in each group ($\alpha_i = \alpha_j = +\frac{1}{2}$, $\beta = 0$, $\gamma = +\frac{1}{2}$).

c. Centroid. On fusion the fused individuals are replaced by the co-ordinates of its centroid which is the sum or mean sum of the individuals forming the group.

i. Squared Euclidean Distance. If the co-ordinate of the centroid of (i) is X; then the centroid of the new fused group (K) will be $(N_iX_i + N_jX_j/N_k)$. Then the difference measure d_{hk} for groups (h) and (K) will be $d_{hK} = \{X_h - (N_iX_i + N_jX_j/N_K\}$. The centroids of groups (i), (j) and (h) are denoted by $X_{(i, j, h)}$ and N_i, N_j, and N_K are the number of individuals in groups (i), (j) and (K). The strategy for this measure is

$$\alpha_i = N_i/N_K,\ \alpha_j = N_j/N_K,\ \beta = -\alpha_i\alpha_j \text{ and } \gamma = 0.$$

ii. Correlation coefficient. For qualitative data, the Pearson ϕ coefficient is usually used and for quantitative data the product-moment coefficient is used. This is = combinatorial, but requires two equations for calculation so it cannot be calculated by the single linear equation.

iii. Non-Metric Coefficient. For binary data, the complement of Czekanowski coefficient is normal. For quantitative data where X_{1j} and X_{2j} are the number of individuals in the jth species in site one and two for site classification

$$\frac{\Sigma|X_{1j} - X_{2j}|}{\Sigma|X_{1j} + X_{2j}|}$$

could be used (Lance and Williams, 1967a).

d. Median. This strategy is similar to squared Euclidean measure except N_i is put equal to N_j which results in

$$\alpha_i = \alpha_j = +\tfrac{1}{2},\ \beta = -\tfrac{1}{4},\ \gamma = 0.$$

e. Group-average. This defines the intergroup distance as the mean of all the between-group interindividual distances (Stephenson et al., 1972). If there are N_i members in group i and N_j members in group j so that in the fused group K there are N_K members, then

$$\alpha_i = N_i/N_K,\ \alpha_j = N_j/N_K,\ \beta = \gamma = 0.$$

f. Incremental sum of squares. The numerical model is Euclidean; the decision-function is the increase on fusion of the sum of the squares of the distance between the individuals and their group centroids (Williams, Clifford and Lance, 1971). Its combinatorial properties are not known.

g. The flexible strategy of Lance and Williams (1967b) is derived from the linear model ($\alpha_i + \alpha_j + \beta = 1$, $\alpha_i = \alpha_j$, β F ω, $\gamma = 0$). It is compatible for Euclidean distance and derives the flexibility from its space distorting properties which will be discussed below.

h. The information statistic strategy is an (i, j, k) measure derived from the information content before and after fusion by the relationship $\Delta I_{(i, j, k)} = I_k - I_i - I_j$. It is a non-combinatorial strategy.

In clustering strategies where an element placed in a group has no effect on its original position in the space, it is said to be <u>space conserving</u>. This property is possessed by any classificatory strategy which uses Euclidean distance between group centroids as its intergroup measure and for AHS of Euclidean systems using centroid fusion strategy. In contrast, the more intensely clustering AHS, the groups appear to recede from each other as they grow. These strategies are called <u>Space Dilating</u>. The receding of groups is group size dependent. This dependence on group size may take one of two forms. It may be <u>asymptotic</u>, so that once the group has attained a modest size further accretions make little difference; or it may be <u>indefinite</u>, so that

every accretion makes the group substantially more remote and therefore substantially more difficult to join (Williams, 1971). Individual/group measures may differ in this respect from the group/group measures. For example, in flexible sorting both measures are asymptotic; in the incremental sum of squares the individual/group measure is asymptotic, the group/group measure indefinite; for the commonest information statistic strategy both are indefinite (Williams et al., 1971).

Choice of Strategies In agglomerative hierarchical clustering strategies, one has to make a choice of the interindividual measure to employ and then the fusion strategy used in building the hierarchy. The choice of interindividual measures is not difficult because it usually is dependent on the nature of the data (Williams, 1971). Highly skewed binary data obtained from presence-and-absence records would usually require Shannon-type information statistics since it is insensitive to skewness and handles such attributes without difficulty (Williams et al., 1971). Euclidean measures which are unduly sensitive to the rare occurences handle data defined by a small number of continuous variables with no strong outliers (Williams, 1971). In contrast, most information statistics are inefficient in dealing with continuous variable (Williams et al., 1971). If the data is non-negative with few zeros, but with an occasional extreme outlier, the Canberra metric is indicated (Williams, 1971). In cases where the data have no striking pecularities, the choice of measure is of relatively little importance, with mixed data it has been observed that information statistics and incremental sum of squares tend to produce nearly identical classifications (Williams, 1971).

The choice of the fusion strategy is much more difficult than the choice of an interindividual measure because there are few hard guidelines to follow. An investigator may have no idea as to the

clustering character of his data or he may wish for greater accuracy at the expense of clearly separate clusters. In such a case, the weakly clustering strategies are the most desirable within the AHS. The appropriate strategies would be nearest-neighbor, group average or centroid strategies. Williams et al. (1971), feel that the group average strategy has some disadvantages in that the system is indefinite. If an attribute is highly skewed with a small number of outlying values, the variance will rise abruptly on fusion with the first of these tending to leave further outlyers stringing along unfused. It also measures identical groups with zero variance so a group of identical individuals acts as an individual, thus very similar individuals forming a group may be vulnerable late in the analysis to capture of an outlying element better placed elsewhere.

The more intensely clustering programs are more appropriate for those workers whose data is either largely continuous and he wishes to chop them into groups as efficiently as possible, or his data may be highly heterogeneous requiring a space dilating strategy. As mentioned earlier, intensely clustering strategies share the common property of increasing the intergroup similarity as group size increases resulting in possible non-conformist groups whose member share only the property that they are unlike everything else including themselves (Lance and Williams, 1967b). A way of relieving this problem is to develop a method for reallocating individuals after the hierarchy is formed to better allocate individuals from a non-conformist group or individuals that were misclassified due to the inherent tendency in AHS for slight misclassification.

The space dilating strategies discussed by Williams et al. (1971) include information statistic, incremental sum of squares and flexibility. Information statistic as mentioned earlier has a strong tendency to produce a non-conformist group, the position of a group in the hierarchy will be strongly dependent on the size of the group and no effective reallocation procedure is known.

Incremental sum of squares has little tendency to produce non-conformist groups although the position of a group in the hierarchy is strongly dependent on the size of the group. Reallocation has been shown to be simple and effective. Flexible strategy does not tend to form the non-conformist group, nor have a marked size dependence on groups entering the hierarchy.

Use of AHS in Benthic Community Analysis As mentioned earlier, agglomerative hierarchical strategies have been the most popular for studying community structure of the macrobenthos. Stephenson and Field have both published several papers where these methods have been used with good results. Although there are no rigid guidelines on which clustering strategy should be used for an individual set of data, the work that has been done to date seems to indicate that one or two strategies are most useful. The habitats that have been analyzed with clustering techniques have been quite diverse ranging from tropical benthos, intertidal, semi-tropical and temporate estuarine, yet the same strategies seem to prevail.

The interindividual measure used most commonly is the complement of Czekanowski or Bray-Curtis similarity index $C_z = 2W/(A+B)$ where A is the sum of the measures of all species in one sample, B is the similar sum for the second sample and W is the sum of the lesser measures of each species for the two samples being compared. This measure was used by Field, 1968 and 1971; Field and McFarland, 1968; Stephenson and Williams, 1971; Stephenson, Williams and Cook, 1972; and Stephenson, et al., 1974. Another interindividual measure was used several times, the Canberra Metric Dissimilarity $d_{1,2} = \frac{1}{N} \sum_{i}^{N} \frac{|X_{1i} - X_{2i}|}{(X_{1i} + X_{2i})}$ where X_{1i} and X_{2i} are the importance values of the ith attribute of the two individuals. That measure was used by Stephenson, Williams and Cook (1972) and Boesch (1973).

The fusion strategies that found the greatest success were the group average and the flexible with $\beta = -0.25$. Field (1968, 1971) used group average only as did Stephenson et al. (1974). Boesch (1973), Stephenson, Williams and Lance (1970), Stephenson, Williams and Cook (1972) used both flexible with $\beta = -0.25$ and group average. Boesch found almost identical results between flexible and group average for the collection site analysis but felt the flexible sorting was much superior to group average for the species analysis. Stephenson and Williams (1971) tried two different information statistic clustering strategies and in both cases found fragmentation of aberrantly rich sites and haphazard nonsense fusion between poorer sites. They resorted to the flexible sorting using $\beta = -0.25$ with satisfactory results.

DIVISIVE HIERARCHICAL STRATEGIES (DHS)

In describing the DHS and AHS earlier, it seems that AHS are fraught with problems that are not encountered with DHS. A natural question would be, why haven't the AHS died out completely in preference for the DHS? The reason is that most practicing DHS programs are based on monothetic division. In contrast, all AHS are polythetic systems based on similarity or dissimilarity. Although the monothetic classifications are simple and fast, they are easy to misclassify. Suppose two groups, X and Y, to be separated monothetically on an attribute possessed by X and lacking in Y. Also suppose individual B possesses the attribute which clusters it with X but more closely resembles the members of Y. At a later stage B will be separated from the main division stemming from X but it will not be able to gain access to the Y side. Because of this, it is characteristic of divisive monothetic systems to produce unduly large numbers of fragmentary groups (Williams, 1971). The polythetic divisive strategies would not do this since it measures overall similarity. They would therefore seem the ideal hierarchical strategy, however,

existing programs require a very small population size and the new program of Wallace and Boulton (1968) are untested on field problems (Williams, 1971).

Monothetic Divisive Two strategies which are based on monothetic division are Association Analysis and Information Analysis. In Association Analysis, the critical attribute is determined by an Index of Association (I_A). The Index of Association in this case is a group measure which determines which should be the next group to be divided or when division should stop. In association analysis of sampling sites, I_A would be calculated for every possible pair of species and the values of I_A entered in an association matrix. The elements of the association matrix are then summed by columns or rows and the critical species to divide the sites on is that having the greatest value of ΣI_A (Pielou, 1969). The index value I_A has been defined in three ways by Williams and Lambert (1959) as χ^2, χ^2 corrected and $\sqrt{\chi^2}/N$.

The information statistic measure proposed by Lance and Williams (1968) calculates the information content, I, of a population as $I = SN \text{ Log } N - \sum_{j=1} \{a_j \text{ Log } a_j + (N-a_j \text{ Log } (N-a_j)\}$ where N site are defined by S species such that the jth species is present in a_j individuals. If a population of sampling sites is divided into two groups (g) and (h), then the information fall would be defined as $\Delta I (gh, i) = I_i - I_g - I_h$. The population is divided at each stage by the attributes for which ΔI is maximum.

Polythetic Divisive As mentioned earlier, divisive-polythetic strategies are more complex and much more time consuming to calculate than divisive monothetic methods. The methods of Edwards and Cavalli-Sforza (1965) have the attributes represented by a single point in a S-dimensional space. The group measure consists of the sum of the

squared distances of all points to the group centroid (Pielou, 1969). The division of a group is based on reducing the within-group sum of squares to a minimum which results in a maximum between-group sum of squares. The method is summarized by Pielou (1969). A population represented by 41 points would take 54,000 years to complete using a computer with access time of 5μ seconds (Gower, 1967). As a result, this method has not found much popularity with benthic ecologists.

A more complicated divisive-polythetic strategy is that proposed by Wallace and Boulton (1968) and Boulton and Wallace (1970) based on information measure. Their program instead of basing group comparison on a measure of similarity or dissimilarity is based on the probability of a member of a group possessing certain measured attributes. The mathematics involved in the strategy are far too complex to discuss in this paper. No mention of computation time is made and since there is no record of anyone using this program in a large study, it remains to be seen whether it would be appropriate for benthic studies.

Use of DHS in Benthic Community Analysis The use of divisive programs in benthic studies has been rather sparse. Because of its greater computational speed over agglomerative and divisive polythetic strategies, divisive monothetic strategies are appropriate with studies containing a great number of stations and/or species.

Moore (1973) employed both association analysis and information analysis in analyzing the fauna of kelp holdfasts. For association analysis, he used $\sqrt{\chi^2}/N$ as the association parameter. The program was "flagged" when no individual χ^2 exceeded 3.84. Although there was some slight misclassification with both strategies, he felt the information analysis gave the best results, but that both strategies were quite satisfactory considering the size of the study.

Stephenson et al. (1970) used information analysis for their divisive strategy. Again, as with Moore, the number of species was exceedingly large, as well as the number of collection sites, so the monothetic approach was the only practical method. After site groups were established using the divisive classification, an agglomerative/ polythetic method was used for the species classification using the site groups, not the individual sites, as attributes.

NON-HIERARCHICAL STRATEGIES

Williams (1971) says of non-hierarchical systems: "for applications in which homogenity of groups is of prime importance, the non-hierarchical strategies are in principle very attractive; unhappily their current state of development lags far behind that of their hierarchical counterparts which at their best are more flexible, provide a wider range of facilities are numerically better understood, and are computationally faster." Perhaps because of these disadvantages, the non-hierarchical systems have received little interest in benthic studies.

The non-hierarchical strategies are divided in two types by Williams (1971). The first type is serially optimized, that is, a group is defined and removed from the population. The remaining individuals are examined and a second group removed; this continues until the entire population is accounted for. The second type is simultaneously optimized where the population is partitioned some way into groups and these are optimized by a repetitive process.

Initiation of a non-hierarchical strategy begins with the calculation if interindividual measures much like the AHS. Later allocation to groups is primarily an individual/group measure. To determine if an individual should be added to a particular group, there have been several sorting strategies proposed most of which Lance and Williams

(1967c) feel are mathematically or computationally unsatisfactory. Of the serially optimized methods discussed, only that of Goodall (1966) was considered mathematically sound.

Goodall's method is based on a probabilistic similarity index used for interindividual measures. Clusters are then formed of all individuals whose maximum dissimilarity is less than a maximum based on a specified significance level. Of the various clusters formed, the largest is removed and the process continued with the residue. The repeated computation of the similarity matrices which is an integral part of the system makes computation time quite lengthy and therefore not suitable for problems of a very large size.

The simultaneously optimized strategies reviewed by Lance and Williams do not suffer from the mathematical inconsistencies of the serial group. Their favorite is MacQueen's (1966) K-mean system. The population is arbitrarily partitioned into K groups and the mean of their Euclidean distance is calculated. As new members are added to the group, the new means are calculated. As a result, the K mean shifts as allocation proceeds. If two group means come closer than a predetermined value, they are fused reducing the number of K groups. Since an upper limit is fixed on allocating an individual to a group, it is possible that an individual cannot be fused so a new nucleus is formed and K increases.

Lance and Williams do not discuss the recurrent group analysis of Fager and McGowan (1963), but it would be considered a serially optimized strategy. The strategy based on interspecific affinities using $J/(NaNb)^{\frac{1}{2}}$ as the interindividual measure. In this index, J is the number of joint occurrences, N_a is the total number of occurrences of species A, N_b is the total number of occurrences of species B. A choice of a significant affinity is chosen and the species are allocated into the largest groups where all species

have affinities equal to or greater than the chosen significance level. The procedure was originally used by Fager and McGowan for zooplankton, but in more recent time, has been used in benthic surveys by Jones (1969), Lie and Kelly (1970), and Boesch (1971).

ADDITIONAL CONSIDERATIONS

Transformations If the range in the numbers of species is very wide, as it often is in marine surveys, the transformation $\ln (N_j+1)$ may be used, where N_j is the number of individuals of the jth species. If the counts are small, they may be distributed as Poisson variates where $\sqrt{N_j}$ transformation will normally be the distribution (Stephenson and Williams, 1971).

Untransformed data was used in the study of Stephenson and Williams (1971) and Stephenson et al. (1970). In the study of the LA Bight, Stephenson et al. (1974) used the cube root transformation.

Standardization of the interindividual dissimilarity measures is also carried out by some workers. Stephenson et al. (1974) used unstandardized measures for the site-groups but standardized by totals for the species-groups. Boesch (1973) used double standardization for site and species groupings.

Double Zero Matches The occurrence of double-zero matches in marine studies where species numbers are usually quite large is common. They are normally dropped in benthic studies but problems can arise whether they are included towards similarity or not. If they are included and there are many of them, the analysis will be dominated by groups having only zero in common (Williams, 1971). If they are ignored and there are some impoverished sites or very heterogeneous data, a hodge-podge of unfusable sites may result or the fusing with any group that contains a species in common (Fields, 1969).

A more difficult problem ecologically is how to handle 0-1 up to perhaps 0-5 matches. Using the Canberra metric dissimilarity coefficient, a 1-0 match has the same dissimilarity as 100-0 match. Boesch (1973) to offset this, suggests substitution of a small constant for zero which he set at 0.001. A number so small has little effect even at the 1-0 level and can't be considered much improvement. Further improvement in this area will hopefully be forthcoming.

Chaining This is a problem in the hierarchical strategies only. Some weakly clustering or space conserving strategies appear on formation of a group to move nearer to some or all the elements. This is not true of all space conserving strategies and has been labelled space-contracting by Lance and Williams (1967b). With such a strategy, the chance of an individual element adding to a pre-existing group is greater than an individual forming the nucleus of a new group. This system is said to be chaining, that is, the tendency of a group to add single individuals or groups much smaller than itself, rather than fusion with groups of comparable size (Williams, Lambert and Lance, 1966). The flexible sorting strategies used in AHS with positive values for β have strong tendencies for chaining. It is felt by Williams et al. (1966) that the more symmetrical the hierarchy, the better the clustering techniques.

Choice of Strategies An investigator deciding to use cluster techniques on his data has wide range of strategies from which to make his choice. Oftentimes, the decision is made for him due to pragmatic reasons such as program availability, computer time, etc. If he has several programs to choose from, his first decision would probably be between a hierarchical and non-hierarchical strategy depending on whether he wanted the cluster as homogeneous and error free as possible or was willing to sacrifice minor misclassification

for more rigid, well-separated groups. Lance and Williams (1967c) suggest a double analysis using the hierarchical system initially, at a subdivision within the hierarchy the data would be analyzed non-hierarchically using a simultaneously optimized strategy such as the K-mean. Once the investigator has made the decision between non-hierarchical or hierarchical, he is mostly on his own.

If the user decides on a hierarchical strategy, the size of the study (i.e., the number of collection sites or number of species collected) may be taken into account. In the case of Moore (1973) where 387 species and 72 sampling sites were involved in the analysis, he felt from a computational standpoint that the divisive monothetic strategy was the only practical choice. Stephenson et al. (1970) analyzed 387 sites and 355 species by first eliminating the rare species (those occurring six times or less). After reduction, they were left with 358 sites and 51 species which were analyzed using a divisive monothetic strategy. The resulting site-groups, eight in all, were then used with an agglomerative strategy for the species-groups.

The methods used in the examples above might be appropriate if the study were very large. However, if the study is not as large as those mentioned above and computer time is not a primary consideration, one of the agglomerative strategies would be the most advisable.

REFERENCES

Bergan, T. 1971. Survey of numerical techniques for grouping. Bacteriological Reviews 35:379-389.

Boesch, D. F. 1971. Distribution and structure of benthic communities in a gradient estuary. Ph.D. Dissertation, College of William and Mary, Virginia.

Boesch, D. F. 1973. Classification and community structure of macrobenthos in the Hampton Roads Area, Virginia. Marine Biology 21:226-244.

Boulton, D. M. and C. S. Wallace. 1970. A program for numerical classification. The Computer J. 13:63-69.

Edwards, A. W. F. and L. L. Cavalli-Sforz. 1965. A method for cluster analysis. Biometrics 21:362-375.

Fager, E. W. and J. A. McGowan. 1963. Zooplankton species-groups in the North Pacific. Science 140:453-460.

Field, J. G. 1968. The use of numerical methods to determine benthic distribution patterns from dredgings in False Bay. Tras, Royal Soc. South Africa 39:183-200.

Field, J. G. 1971. A numerical analysis of changes in the soft-bottom fauna along a transect across False Bay, South Africa. J. Experimental Marine Biology and Ecology 7:215-253.

Field, J. G. and G. McFarlane. 1968. Numerical methods in marine ecology. 1. A quantitative "similarity" analysis of rocky shore samples in False Bay, South Africa. Zoologica Africana 3:119-137.

Field, J. G. 1969. The use of the information statistic in the numerical classification of heterogeneous systems. J. Ecology 57:565-569.

Goodall, D. W. 1966. A similarity index based on probability. Biometrics 22:882-907.

Goodman, L. A. and W. H. Kruskal. 1959. Measures of association for cross-classification. J. American Statistical Association 54:123-163.

Gower, J. C. 1967. A comparison of some methods of cluster analysis. Biometrics 23:623-637.

Jones, G. F. 1969. The benthic macrofauna of the mainland shelf of Southern California. Allan Hancock Monograph in Marine Biology 4. 219 pp.

Lance, G. N. and W. T. Williams. 1967a. Computer programs for hierarchical polythetic classification. The Computer J. 9:60-64.

Lance, G. N. and W. T. Williams. 1967b. A general theory of classification sorting strategies. I. Hierarchical Systems. The Computer J. 9:373-380.

Lance, G. N. and W. T. Williams. 1967c. A general theory of classificatory sorting strategies. II. Clustering Systems. The Computer J. 10:271-277.

Lance. G. N. and W. T. Williams. 1968. Note on a new information statistic classificatory program. The Computer J. 11:195.

Lie, U. and J. C. Kelley. 1970. Benthic in fauna communities off the coast of Washington and in Puget Sound: Identification and distribution of the communities. J. Fisheries Research Board of Canada 27:621-651.

MacQueen, J. B. 1966. Some methods for classification and analysis of multivariate observations. Western Management Science Institute, Univ. of Calif., Working Paper #96.

Moore, P. G. 1973. The kelp fauna of Northeast Britain. II. Multivariate classification: Turbidity as an ecological factor. J. Experimental Marine Biology and Ecology 13:127-163.

Pielou, E. C. 1969. An introduction to mathematical ecology. Wiley-Interscience. New York. 286 pp.

Roback, S. S., J. Cairns and R. L. Kaesler. 1969. Cluster analysis of occurrence and distribution of insect species in a portion of the Potomac River. Hydrobiologia 34:484-502.

Sneath, P. H. A. and R. R. Sokal. 1973. Numerical taxonomy. W. H. Freeman and Co. San Francisco. 573 pp.

Sokal, R. R. and P. H. A. Sneath. 1963. Principles of numerical taxonomy. W. H. Freeman and Co. San Francisco. 359 pp.

Stephenson, W., W. T. Williams and G. N. Lance. 1970. The macrobenthos of Moreton Bay. Ecological Monographs 40:459-494.

Stephenson, W. and W. T. Williams. 1971. A study of the benthos of soft bottoms, Sek Harbour, New Guinea, using numerical analysis. Australian J. Marine and Freshwater Research 22:11-34.

Stephenson, W., W. T. Williams and S. D. Cook. 1972. Computer analysis of Petersen's original data on bottom communities. Ecological Monographs 42:387-415.

Stephenson, W., R. W. Smith, T. S. Sarason, C. R. Greene and D. A. Hotchkiss. 1974. Soft bottom benthos from an area of heavy waste discharge, Palos Verdes, California. I. Hierarchical classification of data from 87 sites. Submitted to J. Experimental Biology and Marine Ecology.

Wallace, C. S. and D. M. Boulton. 1968. An information measure for classification. The Computer J. 11:185-194.

Williams, W. T. 1971. Principles of clustering. In: Annual Review of Ecology and Systematics. 2:303-326.

Williams, W. T. and J. M. Lambert. 1961. Multivariate methods in plant ecology. III. Inverse association analysis. J. Ecology 49:717-729.

Williams, W. T., J. M. Lambert and G. N. Lance. 1966. Multivariate methods in plant ecology. V. Similarity analyses and information analysis. J. Ecology 54:427-446.

This intern report was read and accepted by a staff member at:

Agency: Environmental Protection Agency
Northwest Pacific Environmental Research Laboratory
Address: 200 S.W. 35th Street
Corvallis, Oregon 97365

This report was completed by a WICHE intern. This intern's project was part of the Resources Development Internship Program administered by the Western Interstate Commission for Higher Education (WICHE).

The purpose of the internship program is to bring organizations involved in community and economic development, environmental problems and the humanities together with institutuions of higher education and their students in the West for the benefit of all.

For these organizations, the intern program provides the problem-solving talents of student manpower while making the resources of universities and colleges more available. For institutions of higher education, the program provides relevant field education for their students while building their capacity for problem-solving.

WICHE is an organization in the West uniquely suited for sponsoring such a program. It is an interstate agency formed by the thirteen western states for the specific purpose of relating the resources of higher education to the needs of western citizens. WICHE has been concerned with a broad range of community needs in the West for some time, insofar as they bear directly on the well-being of western peoples and the future of higher education in the West. WICHE feels that the internship program is one method for meeting its obligations within the thirteen western states. In its efforts to achieve these objectives, WICHE appreciates having received the generous support and assistance of the Economic Development Administration; the Jessie Smith Noyes Foundation; the National Endowment for the Humanities; the National Science Foundation; the Division of Education of HEW; and of innumerable local leaders and community organizations, including the agency that sponsored this intern project.

3402-68214520000045600:
1C:1174:GD:WICHE:2H179

For further information, write Bob Hullinghorst, Director, Resources Development Internship Program, WICHE, Drawer 'P', Boulder, Colorado 80302, (303) 443-6144.